Published in final edited form as:

Biometrics. 2010 June ; 66(2): 630–635. doi:10.1111/j.1541-0420.2009.01298.x.

Statistical metrics for quality assessment of high density tiling array data

Hui Tang and Terry M. Therneau

Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA

Hui Tang: h.tang@mayo.edu

Summary

High density tiling arrays are designed to blanket an entire genomic region of interest using tiled oligonucleotides at very high resolution and are widely used in various biological applications. Experiments are usually conducted in multiple stages, in which unwanted technical variations may be introduced. As tiling arrays become more popular and are adopted by many research labs, it is pressing to develop quality control tools as what were done for expression microarrays. We propose a set of statistical quality metrics analogous to those in expression microarrays with application to tiling array data. We also develop a method to estimate the significance level of an observed quality measurement using randomization tests. These methods have been applied to multiple real data sets, including three independent ChIP-chip experiments and one transcriptome mapping study, and they have successfully identified good quality chips as well as outliers in each study.

Keywords

Quality control; Randomization test; Robust linear models; Tiling arrays

1. Introduction

In the last few decades, microarray technology has evolved from covering only annotated transcripts to the so-called “tiling arrays” which blanket a whole genome using evenly-spaced probes at very high resolution. The number of probes planted on the arrays has increased from several thousands to several millions, a trend quite similar to the number of transistors on a silicon chip in the semiconductor industry (Mockler and Ecker, 2005). For example, the Affymetrix Human Genome U133 microarray contains almost 45,000 probe sets representing over 33,000 genes. The GeneChip Chromosome 21/22 1.0 tiling array set consists of over one million oligonucleotide probes to cover these two chromosomes at 35-base pair (bp) resolution (genomic resolution is measured from the central position of adjacent oligos). And the Human Tiling 2.0R array set contains approximately 45 million probes to interrogate the human genome at 35-bp resolution. As the techniques advance, it is expected that more probes will be planted on tiling arrays with greater and greater resolutions.

Tiling arrays have broad applications in genome-wide mapping of transcription factor binding sites (ChIP-chip experiments), transcriptome mapping, DNA methylation site mapping, high density arrayCGH for comparative genomic hybridization, and many other studies (Liu, 2007). Given the large volume of data acquired in a single experiment, they

have posed challenging problems for statisticians. The first issue is quality control and assessment. There are many sources of variability in the experiment, including real biological differences (e.g. animal to animal, experimental conditions, etc) and technical noise introduced during multiple experimental stages such as chromatin sonication, RNA isolation and handling, amplification of DNA fragment mixtures, labeling and hybridization conditions, scanner characteristics, etc. A single chip costs several thousand dollars, which makes experiments with multiple conditions and replications quite expensive. It is crucial to have quality check procedures before an exhaustive experiment is conducted to make sure the samples are of good quality and that unwanted technical noise is reduced to the lowest level possible.

The statistical and computational literature has been focusing on various topics of tiling array design and analysis, such as array design optimization (Graf et al, 2007; Bertone et al., 2006; Emanuelsson et al, 2007), pre-processing and normalization methods (Royce et al., 2007), and interval prediction programs (Johnson et al, 2006; Ghosh et al, 2007), but none has had a thorough discussion of quality check and assessment issues. According to Nadon and Shoemaker (2002), there are two types of measurement error: random and systematic. Random error is “a measure of uncertainty” and is “inevitable”. The random error cannot be eliminated and can be estimated from observed data by using statistical tests. Systematic errors are “biases, which result in a constant tendency to over- or under-estimate true values and may affect all expression values on an array equally”. These authors suggested the use of multiple replications (≥ 3 per group or experimental condition) from the same DNA sample to help accurately estimate random error and reduce systematic error.

Brettschneider et al (2008) and Heber and Sick (2006) offered a comprehensive review of statistical tools for quality assessment of expression microarrays. These authors proposed the use of parallel plots of RLE (relative log expression) and NUSE (normalized unscaled standard error) to visualize outliers in a group of microarrays. They suggested a “rule of thumb” such as “chips with median of NUSE bigger than 1.05 is alarming”, or to simply look for trends, patterns and outliers inconsistent with others. One question that remains unanswered is how to assess the significance levels of these proposed scores.

In this paper we will extend the underlying framework of these quality metrics and make them applicable in high density tiling array data. These metrics are estimated from fitting a robust linear model to the logarithm of the preprocessed probe intensities for a particular probe set. For each chip in a group, a significance level for an observed quality measurement will be estimated based on a randomization test (Manly, 2007). Chips with extreme significance levels indicate that technical or experimental noise may have been introduced during sample preparation and array production procedures, and in downstream analysis these arrays should be used with caution.

2. The NUSE and RLE statistics in tiling arrays

2.1 Special characteristics of tiling array data

Compared to gene expression microarrays, high-density tiling arrays have several differences. First, in expression microarrays, cRNA samples are labeled and hybridized on arrays, while in tiling array experiments, it could be either cDNA (ChIP-chip) or cRNA (transcripts mapping) samples. Second, in expression microarrays, except for custom-made arrays with only a few hundred genes such as the Illumina DASL array (Illumina Inc., 2005), it is reasonable to assume that the majority of genes are not biologically affected, which is one of the key assumptions made in Brettschneider et al (2008). In tiling arrays, this assumption holds only under certain circumstances. For ChIP-chip experiments used to study genome-wide transcription factor binding sites, we expect binding signals only in

regions where the protein and DNA interacts, therefore it is valid to assume that most genomic regions do not have binding signals. For transcripts mapping studies, the assumption does not hold, as we expect that changes in RNA levels could occur in many genomic intervals as observed in previous studies (Bertone et al., 2004; Cheng et al, 2005). Finally, tiling arrays do not have the concept of probe sets, oligos are not clustered into regions but are tiled uniformly across the non-repetitive regions.

2.2 Relative log expression and a robust linear model

Relative log expression (RLE) and normalized unscaled standard error (NUSE) were statistical tools developed for quality assessment in expression microarrays (Brettschneider et al, 2008; Heber and Sick, 2006). RLE computes how much the measurement of the expression of a particular probe set in a chip deviates from that in other chips. As pointed out by Brettschneider et al (2008), under the two assumptions that (1) “the majority of genes are not biologically effected”, and (2) “the number of up regulated genes $\approx$ the number of down regulated genes”, the interquartile range (IQR) of RLE and the difference of the median of RLE from zero are measures of technical variability in that chip. NUSE is based on a robust linear model fit to each probe set. Probes within a set are supposed to be inherently consistent in terms of expression. NUSE measures the relative precision of expression estimates. For a good quality expression array, the median of NUSE should be around one and the IQR of NUSE should be small and consistent with other replication chips.

We propose an innovative way to apply RLE and NUSE to the quality assessment of high density tiling arrays. A study-dependent window size is used to partition genomic regions into pseudo probe sets. The rationale is to make the length of the set as close to the underlying biological dimension as possible so that the hybridization signals within a pseudo probe set are inherently consistent. For example, in a transcription factor binding sites mapping study, we typically use a window size of 1 kb, because we believe that most protein-DNA interactions occur within that range; while in a novel transcripts study in which people are interested in finding long transcripts (say, > 500-bp), we use a window size of 500-bp instead.

To calculate RLE for n tiling arrays, $i = 1, \dots, n$, we first construct a reference chip $n + 1$, where the value of each probe is the median of the other n chips. For a probe set k , we first summarize the probe intensities using median polish (Tukey, 1977), then RLE for the probe set k on chip i is the log ratio of its intensity summary on chip i and that on the reference chip: $\log(I_{ik}/I_{n+1,k})$. RLE can be used to assess the quality of a tiling array by plotting parallel boxplots for all or a sample (when the number of probe sets is huge) of probe sets on each array. The IQR of RLE is a measure of difference from the reference chip, which is an indicator of either biological or technical variability or both on that array. Under the assumption “the majority of probe sets do not exhibit signals” (e.g. in ChIP-chip experiments), the IQR of RLE becomes a measure of technical variability on that chip. Even if the assumption does not hold, the IQR of RLE can still detect technical variability that is larger than biological variability. The median of RLE is a measure of bias, and we expect to see that it is close to zero in a good quality chip under the assumption “the number of probe sets having positive signals is approximately equal to the number of probe sets having negative signals” (Note: the “positive” and “negative” signals are with respect to the reference chip).

To calculate NUSE for n arrays, consider fitting the following model for each probe set:

$$\log(I_{ij}) = u_i + a_j + e_{ij} \quad (1)$$

where u_i is a chip effect, $i = 1, \dots, n$, n is the number of chips; a_j is a probe effect, $j = 1, \dots, p$, p is the number of probes in a fixed probe set; and e_{ij} is the probe-chip wise error assumed to be i.i.d. centered with standard deviation σ .

Most expression microarrays are designed to have about 20 probes in each probe set. For tiling arrays, the number of probes in each probe set ranges from 16-60 based on the partition method mentioned above. These are reasonable ranges for fitting the model in equation (1). For each probe set, the model is fitted robustly using iteratively re-weighted least squares (Venables and Ripley, 2002). The analysis will generate the intensity estimates $\hat{u}_i$ for the probe set on chip i , residuals r_{ij} and weights w_{ij} attached to probe j on chip i . NUSE is actually the relative precision of each of $\hat{u}_i$ defined as

$$NUSE(\hat{u}_i) = \frac{se(\hat{u}_i)}{\text{median}_i(se(\hat{u}_i))} = \frac{1 / \sqrt{W_i}}{\text{median}_i(1 / \sqrt{W_i})} \quad (2)$$

where $W_i = \sum_j w_{ij}$ is the total probe weight for the probe set on chip i . For a good quality tiling array, the median of NUSE is close to its expected value one and the IQR of NUSE should be small and consistent with other replication chips.

These statistical metrics for quality assessment of tiling arrays are summaries over a genomic region on how noisy the samples are. Instead of plotting intensities for millions of probes, wondering what are the real biological effects and what are technical noises, people now are able to visualize and summarize the experimental noise by using these quality metrics. It is convenient to identify outlier chips and spot trends across parallel boxplots of these quality metrics.

In ChIP-chip studies, the following assumptions hold: (1) the majority of probe sets do not exhibit signals; (2) the number of probe sets having positive signals (relative to the reference chip) is approximately equal to the number of probe sets having negative signals. Thus we can use both RLE and NUSE, and we will show that NUSE is much more sensitive to detect array differences than RLE in example data sets in Section 3.2.1 and Web Appendix A. For studies where a large portion of the genome exhibits signals, such as transcriptom mapping studies (David et al, 2006), the assumptions for using median and IQR of RLE do not hold, but NUSE is still applicable. An example will be given in Section 3.2.2.

2.3 Estimation of significance levels using randomization tests

Besides visualizing parallel NUSE boxplots, we utilized a formal randomization test procedure to identify outlier chips. The null hypothesis is that there is no outliers, i.e. the error distribution is identical for all chips. The alternative is that there is at least one outlier whose error distribution is different from others. The test statistic is the difference of median(NUSE) from its expected value 1, where NUSE is estimated from the model in equation (1). We generated the null distribution as follows: (i) randomly permute the residuals across chips while preserving the probe set structure; (ii) add permuted residuals to fitted values; and (iii) fit the robust model again using data in (ii) to calculate a new set of NUSE and the difference. Steps (i) to (iii) are repeated many times, we then compared the observed difference with the distribution generated by random permutation.

We have conducted a simulation study to study two properties of the test: (i) when the null is true, i.e. there is no outliers, how often is one wrongly declared (false positive rate)? (ii)

when there is an outlier chip with a different error distribution, what is the chance of failing to claim it (false negative rate or 1-power)? This is more important in the context of this paper as our aim is to identify potential outliers for further inspection. For all simulations, the null distribution of the residuals was a contaminated normal distribution $0.9N(0, 1) + 0.1N(0, 10)$, e.g. some probe failures are expected even in a “good” experiment. Under the null, all errors come from this distribution; under the alternative, one outlier chip's error comes from a $N(0, 2)$. For simulated experiments with $n = 3, 10, 30$ or 100 chips, the type I error of our method was quite conservative for $n = 3$, and only slightly conservative for the other cases (Web Table 1). Under the alternative, the aberrant chip was universally detected. Complete details of the simulation study are in Web Appendix B.

3. Applications to high-density tiling array data

3.1 Data sets

The following four public data sets will be used to illustrate the use of the proposed methods (to save space in the main paper, the second and third data sets are in Web Appendix A): (1) Chromosome-wide mapping of Estrogen Receptor (ER) binding sites in MCF7 cells. This is a ChIP-chip experiment using Affymetrix chromosome 21/22 1.0 array-set to map in-vivo ER binding sites (Carroll et al, 2005); (2) A simulated ChIP-chip experiment using mixtures of human genomic DNA and predefined DNA targets or “spike-ins” at various concentrations hybridized to four tiling array platforms (Johnson et al, 2008). We will use the Affymetrix ENCODE v2 data generated in Struhl lab for illustration; (3) A ChIP-on-Chip experiment for mapping binding sites for two transcription factors (ER α and c-MYC) in MCF7 cells. This is a ChIP-chip experiment using Affymetrix Human Tiling 2.0R platform (7-array set) (Hua et al, 2008); (4) Whole-genome transcripts mapping in normal human bronchial/tracheal epithelial (NHBE) cells. This is a novel transcripts mapping study using Affymetrix Human Tiling 1.0R platform (14-array set) (Perez et al, 2008).

3.2 Results

3.2.1 Chromosome-wide mapping of ER binding sites in MCF7 cells using

Affymetrix chromosome 21/22 1.0 platform—This array set contains A, B and C arrays, each with about 300,000 probe pairs. Three ER+ ChIP samples were hybridized by using MCF7 cells 45 min after ER activation, and three genomic input samples were hybridized by using the MCF7 genomic input DNA.

For illustration purpose, only array type A is plotted. A window size of 1 kb is chosen to partition the genome. Figure 1 shows parallel boxplots of NUSE for a random sample of 2000 probe sets on array type A for the six chips. Note the third control has both larger median(NUSE) and IQR(NUSE), indicating the noise level is higher than other five samples. To estimate a significance level for the difference of the median NUSE from 1, we run 1000 random permutations for these probe sets on all six chips. For each chip, the permuted differences between the median of NUSE and 1 are ranked and compared to the observed value. The third control has an estimated significance level of 0.001, meaning that among 1000 randomly permuted differences, none is bigger than the observed one, indicating that the difference is quite impossible due to chance alone and the third control is an outlier. Our finding agrees with the fact that during the experiment there were not enough DNA left after PCR amplification, so only less than 1 μ g material was added to the third control chip (Johnson et al, 2006).

The outlier is also reflected by the RLE parallel boxplots as in Figure 2. By comparing these two figures, we can see that the RLE is less sensitive in detecting outlier chips compared to

NUSE, which coincides with the finding in expression microarrays (Brettschneider et al, 2008).

3.2.2 Non-coding transcripts mapping study in NHBE cells using Affymetrix Human Tiling 1.0R Arrays—This platform contains 14 arrays covering the human genome, with each array containing over 3.25 million probe pairs (PM and MM). The experiments were performed in triplicates for the normal cell line (NHBE), and duplicates for the other two stressed cells (Hypoxia and NNK).

This was a transcriptome mapping study where people were interested in finding long (at least 800-bp), novel, non-coding transcripts, therefore a window size of 800-bp is used to partition the genome. Again for illustration purpose, only one array type M01R is plotted here. Figure 3 is the parallel boxplots of NUSE for 2000 random probe sets. The third replicate of NHBE is apparently an outlier, which is due to the day effect and protocol change — the first two samples were labeled and hybridized together and the third sample was hybridized at a later date and included a ribosomal removal step in the protocol (via personal communications with the author). The estimated significance levels for all seven chips are 0.001, suggesting that the technical variability was also introduced in the Hypoxia and NNK samples, although they were not detected by just looking at the parallel boxplots.

4. Discussion

In this paper we have developed a set of statistical metrics for quality assessment of high-density tiling arrays analogous to tools used in expression microarray analysis. Successful applications of these metrics have been illustrated using four independent tiling array experiments. Note that by partitioning tiled probes into pseudo sets, a similar probe-set concept to that in expression arrays has been created, and many tools developed in the expression microarray data analysis can be carried over to high density tiling array data. We also studied the effect of different window sizes to these metrics (Web Appendix C) and the result showed that they are quite robust.

Due to the fact that each array platform has its unique probe spacing pattern, the model in equation (1) can not be applied across different array platforms. This creates interesting topics in experimental design. For example, for the spike-in experiment used in Web Appendix A, multiple ChIP-chip experiments have been conducted using four platforms, two concentrations, two conditions (control and Spike-in), in eight different labs. However, one array platform was used exclusively in one lab, which means that we can evaluate the variability due to the concentration effect, the hybridization effect and the scanner effect, but not the lab effect.

Quality control and assessment is one of the most studied topics in manufactured commercial microarrays. One reason is their broad application in biological and medical research, for example, gene expression profiling, transcription regulation study, discovery of new surrogate biomarkers, etc. Another reason is that there are many sources of variation during an experiment. Different protocols might be adopted in different labs based on disease type, tissue type/cell line, treatment type and study aims. It would be highly desirable to ensure the quality of these microarrays for sound and valid scientific findings. Labs with small budgets can start with a less expensive set of arrays, for example, the promoter 1.0R array, or one or two arrays from a larger and more expensive array set such as the Human Tiling 1.0R or 2.0R array sets. By using our method to quickly check the quality of the samples, one can choose the protocol that best minimizes the technical error introduced during experimental procedures and proceed to an exhaustive study with confidence in protocol and sample qualities.

In case-control studies, people often use treatment versus control or untreated patients. Similarly, in ChIP-chip, a treatment array means that that cells are first treated with a drug or a chemical, then cross-linked, sonicated, and immunoprecipitated with a specific antibody. Then the immunoprecipitated DNA samples are amplified, labeled and hybridized to tiling arrays. A control array, however, can mean two different things: one type of control is the “real” control, whose procedures are exactly the same as in the treatment array, except that the cells are untreated. The other type of control is called “input”, which is simply the genomic DNAs (e.g. chromosomes have been sonicated, but not immunoprecipitated). In ChIP-chip study design, even if the aim is to compare treated and untreated samples, getting an input control is highly recommended in order to eliminate the genomic noise. The number of replication arrays is also an important factor to ensure the quality of the experiment and better estimate of random errors. Usually it is recommended to have at least three or four replicates per group as in expression microarrays (Nadon and Shoemaker, 2002; Lee et al, 2000). Systematic evaluation of the tiling array variability and quality control methods are favorable such as the strategies used in the MicroArray Quality Control (MAQC) Project (MAQC Consortium, 2006) and the systematic study of variability for ChIP-chip experiments (Johnson et al, 2008).

The code is written in R and is available from the author.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully thanks Drs. Terence P. Speed and Benjamin M. Bolstad for helpful discussions and insights. This work was supported by Mayo Clinic Cancer Center Grant.

References

- Bertone P, Stolc V, Royce TE, et al. Global identification of human transcribed sequences with genome tiling arrays. *Science*. 2004; 306:2242–2246. [PubMed: 15539566]
- Bertone P, Trifonov V, Rozowsky JS, et al. Design optimization methods for genomic DNA tiling arrays. *Genome Research*. 2006; 16(no. 2):271–281. [PubMed: 16365382]
- Brettschneider J, Collin F, Bolstad BM, Speed TP. Quality assessment for short oligonucleotide microarray data. *Technometrics*. 2008; 50(no.3):241–264.
- Carroll JS, Liu XS, Brodsky AS, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein foxa1. *Cell*. 2005; 122(no.1):33–43. [PubMed: 16009131]
- Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*. 2005; 308:1149–1154. [PubMed: 15790807]
- David L, Huber W, Granovskaia M, et al. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103(no. 14):5320–5325. [PubMed: 16569694]
- Emanuelsson O, Nagalakshmi U, Zheng D, et al. Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Research*. 2007; 17(no. 6):886–897. [PubMed: 17119069]
- Ghosh S, Hirsch HA, Sekinger EA, et al. Differential analysis for high density tiling microarray data. *BMC Bioinformatics*. 2007; 8:359. [PubMed: 17892592]
- Graf S, Nielsen FGG, Kurtz S, et al. Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*. 2007; 23(no.13):i195–204. [PubMed: 17646297]
- Heber S, Sick B. Quality assessment of Affymetrix GeneChip data. *OMICS A Journal of Integrative Biology*. 2006; 10(no.3):358–368. [PubMed: 17069513]

- Hua S, Kallen CB, Dhar R, et al. Genomic analysis of estrogen cascade reveals histone variant H2A.Z associated with breast cancer progression. *Molecular System Biology*. 2008; 4(no. 188):1–14.
- Illumina Inc. DASL assay vs. Direct Hybridization: Understanding the Differences Between Illumina Two Gene Expression Assays. 2005. <http://www.illumina.com/downloads/DASLvsDirectHyb.pdf>
- Johnson WE, Li W, Meyer CA, et al. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103:12457–12462. [PubMed: 16895995]
- Johnson DS, Li W, Gordon DB, et al. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Research*. 2008; 0 gr.7080508.
- Lee MLT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*. 2000; 97:9834–9839. [PubMed: 10963655]
- Liu XS. Getting started in tiling microarray analysis. *PLoS Computational Biology*. 2007; 3(no. 10):e183, 1842–1844.
- Manly, BFJ. Randomization, Bootstrap and Monte Carlo Methods in Biology. Third. Chapman & Hall/CRC; 2007.
- MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*. 2006; 24(no. 9):1151–61.
- Mockler TC, Ecker JR. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*. 2005; 85:1–15. [PubMed: 15607417]
- Nadon R, Shoemaker J. Statistical issues with microarrays: processing and analysis. *Trends in Genetics*. 2002; 18(no.5):265–271. [PubMed: 12047952]
- Perez DS, Hoage TR, Pritchett JR, et al. Long, abundantly expressed non-coding transcripts are altered in cancer. *Human Molecular Genetics*. 2008; 17(no. 5):642–655. [PubMed: 18006640]
- Royce TE, Rozowsky JS, Gerstein MB. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*. 2007; 23(no. 8):988–997. [PubMed: 17387113]
- The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007; 447:799–816. [PubMed: 17571346]
- Tukey, JW. *Exploratory Data Analysis*. Addison-Wesley; Reading, Massachusetts: 1977.
- Venables, WN.; Ripley, BD. *Modern Applied Statistics with S*. Fourth. Springer; 2002.

Figure 1. Parallel boxplots of NUSE for 2000 random probe sets in MCF7 ER regulation study. Labels: “ER”: ER-specific antibody; “INP”: genomic input DNA. Three replicates for each. The third control “INP.3” is an outlier.

Figure 2. Parallel boxplots of RLE for 2000 random probe sets in MCF7 ER regulation study. Labels are the same as Figure 1. The outlier “INP.3” shows a lightly larger IQR compared to the other five chips, which is consistent with NUSE (but less sensitive).

Figure 3. Parallel boxplots of NUSE for 2000 random probe sets for the novel transcripts mapping study in NHBE cells. Labels: “NHBE”: control (three replicates); “NNK” and “Hypoxia” are two stress conditions (two replicates each).